

Confidence Interval for the Mean of a Bounded Random Variable and Its Applications in Point Estimation *

Xinjia Chen

February, 2008

Abstract

In this article, we derive an explicit formula for computing confidence interval for the mean of a bounded random variable. Moreover, we have developed multistage point estimation methods for estimating the mean value with prescribed precision and confidence level based on the proposed confidence interval.

1 Introduction

In many areas of sciences and engineering, it is a frequent problem to estimate the mean of a bounded random variable. Conventional technique for constructing confidence interval relies on the Central Limit Theorem. However, for small and moderate sample size, using normal approximation can lead to serious under-coverage of the mean. In the case of bounded random variables, even the sample size is very large, the error can also be intolerable when the parent distribution is highly skewed toward extremes.

In this article, by applying an inequality obtained by Massart 1990 and Hoeffding's probability inequality, we have derived an explicit formula for interval estimation of the mean in the bounded case. The formula is extremely simple. Moreover, we have proposed multistage estimation methods for estimating the mean value with prescribed precision and confidence level based on the construction of confidence interval.

2 Explicit Formula

Since any random variable X bounded in interval $[a, b]$ (i.e., $\Pr\{a \leq X \leq b\} = 1$) has a linear relation with random variable $Z = \frac{X-a}{b-a}$, it suffices to consider interval estimation for the mean

*The author is currently with Department of Electrical Engineering, Louisiana State University at Baton Rouge, LA 70803, USA, and Department of Electrical Engineering, Southern University and A&M College, Baton Rouge, LA 70813, USA; Email: chenxinjia@gmail.com

of random variable Z on interval $[0, 1]$ (i.e., $\Pr\{0 \leq Z \leq 1\} = 1$) and employ transformation $X = (b - a)Z + a$ to obtain an estimation for the mean of X . The following Theorem 1 provides an easy method for constructing confidence interval for the mean of Z .

Theorem 1 Let $\delta \in (0, 1)$ and $c = \frac{9}{2 \ln \frac{2}{\delta}}$. Let $\Pr\{0 \leq Z \leq 1\} = 1$ and $\mu = \mathbb{E}(Z)$. Let $\bar{Z} = \frac{\sum_{i=1}^n Z_i}{n}$ where n is the sample size and $Z_i, i = 1, \dots, n$ are i.i.d. observations of Z . Define

$$L = \bar{Z} + \frac{3}{4 + nc} \left[1 - 2\bar{Z} - \sqrt{1 + nc\bar{Z}(1 - \bar{Z})} \right],$$

$$U = \bar{Z} + \frac{3}{4 + nc} \left[1 - 2\bar{Z} + \sqrt{1 + nc\bar{Z}(1 - \bar{Z})} \right].$$

Then,

$$\Pr\{L < \mu < U\} \geq 1 - \delta.$$

To prove Theorem 1, we need some preliminary lemmas.

Lemma 1 Let $\alpha = \frac{1}{nc}$. Let $0 \leq t \leq 1$. Then $\epsilon(t) = \frac{3\alpha(1-2t)+3\sqrt{\alpha^2+4\alpha t(1-t)}}{2(1+\alpha)} \geq 0$ satisfies equation

$$\exp\left(-\frac{n\epsilon^2}{2(t + \frac{\epsilon}{3})(1 - t - \frac{\epsilon}{3})}\right) = \frac{\delta}{2} \quad (1)$$

with respect to ϵ .

Proof. Let $q = t + \frac{\epsilon}{3}$ where ϵ satisfies equation (1). Then q satisfies equation $\exp\left(-\frac{9n(q-t)^2}{2q(1-q)}\right) = \frac{\delta}{2}$, which can be simplified as

$$(q - t)^2 + \alpha q(q - 1) = 0 \quad (2)$$

with two real roots $q = \frac{2t + \alpha \pm \sqrt{\alpha^2 + 4\alpha t(1-t)}}{2(1+\alpha)}$. Making use of the relation between ϵ and q , we find the roots of equation (1) as $\epsilon_1 = \frac{3\alpha(1-2t)+3\sqrt{\alpha^2+4\alpha t(1-t)}}{2(1+\alpha)}$ and $\epsilon_2 = \frac{3\alpha(1-2t)-3\sqrt{\alpha^2+4\alpha t(1-t)}}{2(1+\alpha)}$. It can be verified that $|\alpha(1-2t)|^2 \leq \alpha^2 + 4\alpha t(1-t)$, which leads to $\epsilon(t) = \epsilon_1 \geq 0$ and $\epsilon_2 \leq 0$. \square

Lemma 2 Let $t \in (0, 1)$. Then $\epsilon(t)$ is a concave function with respect to t .

Proof. By equation (2), we have $0 < t < q < 1$ and $\frac{dq}{dt} = \frac{2(q-t)}{2(q-t)+\alpha(2q-1)} = \frac{1}{1+\alpha+\frac{\alpha(t-\frac{1}{2})}{q-t}}$. Conse-

quently, $\frac{d\left(\frac{t-\frac{1}{2}}{q-t}\right)}{dt} > 0 \iff (q-t) - (t-\frac{1}{2})\left(\frac{dq}{dt} - 1\right) > 0 \iff q-t > \frac{\alpha(t-\frac{1}{2})(1-2q)}{2(q-t)+\alpha(2q-1)}$. Moreover,

$\frac{d^2\epsilon}{dt^2} = 3\frac{d^2q}{dt^2} = \frac{-3\alpha}{\left[1+\alpha+\frac{\alpha(t-\frac{1}{2})}{q-t}\right]^2} \frac{d\left(\frac{t-\frac{1}{2}}{q-t}\right)}{dt}$. Therefore, to show $\frac{d^2\epsilon(t)}{dt^2} < 0$, it suffices to show inequal-

ity $q-t > \frac{\alpha(t-\frac{1}{2})(1-2q)}{2(q-t)+\alpha(2q-1)}$, which is equivalent to $1 > \frac{\alpha(t-\frac{1}{2})(1-2q)}{2(q-t)^2+\alpha(q-t)(2q-1)}$ since $q-t > 0$. Note that $\frac{\alpha(t-\frac{1}{2})(1-2q)}{2(q-t)^2+\alpha(q-t)(2q-1)} = \frac{\alpha(t-\frac{1}{2})(1-2q)}{2(q-t)^2+2\alpha q(q-1)+\alpha q-\alpha t(2q-1)} = \frac{(t-\frac{1}{2})(1-2q)}{q-t(2q-1)}$ because q satisfies equation

(2). It follows that, to show $\frac{d^2\epsilon(t)}{dt^2} < 0$, it suffices to show inequality $1 > \frac{(t-\frac{1}{2})(1-2q)}{q-t(2q-1)}$. Invoking inequality $0 < t < q < 1$, we can show that $q - t(2q - 1) > 0$, which leads to equivalent relations $1 > \frac{(t-\frac{1}{2})(1-2q)}{q-t(2q-1)} \iff q - t(2q - 1) > (t - \frac{1}{2})(1 - 2q) \iff 0 > -\frac{1}{2}$. The last inequality is trivially true. \square

Lemma 3 Let $\beta = \frac{4}{nc}$. Let $t(z) = z + \frac{3\beta(1-2z)-3\sqrt{\beta^2+4\beta z(1-z)}}{4(1+\beta)}$ where $0 \leq z \leq 1$. Then $z - t(z) = \epsilon(t(z))$ and $t(z) \leq z$.

Proof. Let $p = t + \frac{z-t}{3}$ where t satisfies $z - t = \epsilon(t)$. It follows that $\epsilon(t) = \frac{-3(p-z)}{2}$ and $t + \frac{\epsilon(t)}{3} = p$. By Lemma 1, $\epsilon(t)$ satisfies equation (1), hence p satisfies equation $\exp\left(-\frac{\frac{9}{4}n(p-z)^2}{2p(1-p)}\right) = \frac{\delta}{2}$, which can be simplified as $(p - z)^2 + \beta p(p - 1) = 0$ with two roots $p = \frac{2z+\beta \pm \sqrt{\beta^2+4\beta z(1-z)}}{2(1+\beta)}$. Making use of the relation between p and t , we find the solution of equation $z - t = \epsilon(t)$ with respect to t as $t_1 = z + \frac{3\beta(1-2z)+3\sqrt{\beta^2+4\beta z(1-z)}}{4(1+\beta)}$ and $t_2 = z + \frac{3\beta(1-2z)-3\sqrt{\beta^2+4\beta z(1-z)}}{4(1+\beta)}$. It can be shown that $|\beta(1 - 2z)|^2 \leq \beta^2 + 4\beta z(1 - z)$, which leads to $t_1 \geq z$ and $t_2 \leq z$. So the proof is completed by noting that $t(z) = t_2$. \square

Lemma 4 Let $0 < \mu < 1$ and $0 \leq z \leq 1$. Then $z - \mu \geq \epsilon(\mu)$ if $t(z) \geq \mu$.

Proof. Let $t(z) \geq \mu > 0$. By Lemma 3, we have $z - t(z) \geq 0$ and thus $z - \mu \geq z - t(z) \geq 0$. We claim that $z - \mu > 0$. If this is not true, then $z = \mu$ and $t(z) \geq z > 0$. By Lemma 3, we have $t(z) = z > 0$. On the other hand, $t(z) = z$ results in $z = 0$. Thus we arrive at contradiction $0 > 0$. So we have shown $z - \mu > 0$ and it follows that $0 \leq \frac{z-t(z)}{z-\mu} \leq 1$. We next show that $z - \mu \geq \epsilon(\mu)$. Suppose for the purpose of contradiction that $z - \mu < \epsilon(\mu)$. Then

$$z - t(z) = (z - \mu) \frac{z - t(z)}{z - \mu} < \epsilon(\mu) \frac{z - t(z)}{z - \mu} + \left(1 - \frac{z - t(z)}{z - \mu}\right) \epsilon(z).$$

By Lemma 2, $\epsilon(t)$ is concave with respect to t , hence $\epsilon(\mu) \frac{z-t(z)}{z-\mu} + (1 - \frac{z-t(z)}{z-\mu})\epsilon(z) < \epsilon(t(z))$, which yields $z - t(z) < \epsilon(t(z))$. Recall Lemma 3, $z - t(z) = \epsilon(t(z))$. It follows that $\epsilon(t(z)) < \epsilon(t(z))$, which is a contradiction. \square

We are now in the position to prove Theorem 1. By Theorem 1 of Hoeffding 1963,

$$\Pr\{\bar{Z} \geq \mu + \epsilon\} \leq \left\{ \left(\frac{\mu}{\mu + \epsilon} \right)^{\mu + \epsilon} \left(\frac{1 - \mu}{1 - \mu - \epsilon} \right)^{1 - \mu - \epsilon} \right\}^n \quad \forall \epsilon \in (0, 1 - \mu). \quad (3)$$

By Lemma 1 of Massart 1990,

$$(\mu + \epsilon) \ln \left(\frac{\mu + \epsilon}{\mu} \right) + (1 - \mu - \epsilon) \ln \left(\frac{1 - \mu - \epsilon}{1 - \mu} \right) \geq \frac{\epsilon^2}{2(\mu + \frac{\epsilon}{3})(1 - \mu - \frac{\epsilon}{3})} \quad \forall \epsilon \in (0, 1 - \mu). \quad (4)$$

It follows from (3) and (4) that

$$\Pr\{\bar{Z} \geq \mu + \epsilon\} \leq \exp\left(-\frac{n\epsilon^2}{2(\mu + \frac{\epsilon}{3})(1 - \mu - \frac{\epsilon}{3})}\right) \quad \forall \epsilon > 0. \quad (5)$$

By the definition of $t(\cdot)$, we can verify that $L = t(\bar{Z})$. Thus $\Pr\{L \geq \mu\} = \Pr\{t(\bar{Z}) \geq \mu\}$. Applying Lemma 4, we have $\Pr\{t(\bar{Z}) \geq \mu\} \leq \Pr\{\bar{Z} - \mu \geq \epsilon(\mu)\}$. Hence by (5) and Lemma 1,

$$\Pr\{L \geq \mu\} \leq \Pr\{\bar{Z} - \mu \geq \epsilon(\mu)\} \leq \exp\left(-\frac{n[\epsilon(\mu)]^2}{2(\mu + \frac{\epsilon(\mu)}{3})(1 - \mu - \frac{\epsilon(\mu)}{3})}\right) = \frac{\delta}{2}.$$

Since $\Pr\{L \geq \mu\} \leq \frac{\delta}{2}$ has been shown, applying this conclusion to random variable $1 - Z$, we have $\Pr\{U \leq \mu\} \leq \frac{\delta}{2}$.

Finally, by applying Bonferroni's inequality, we have

$$\begin{aligned} \Pr\{L < \mu < U\} &\geq \Pr\{L < \mu\} + \Pr\{U > \mu\} - 1 \\ &= 1 - \Pr\{L \geq \mu\} + 1 - \Pr\{U \leq \mu\} - 1 \\ &\geq 1 - \frac{\delta}{2} + 1 - \frac{\delta}{2} - 1 = 1 - \delta. \end{aligned}$$

3 Applications in Multistage Point Estimation

We would like to note that the simple interval estimation method described above can be used to construct multistage sampling plans for estimating the mean value of a bounded variable with prescribed precision and confidence level. To illustrate such applications, we shall first present some general results of multistage point estimation based on confidence intervals.

Let X be a random variable parameterized by θ , which is not necessary bounded. Let X_1, X_2, \dots be a sequence of random samples of X . The goal is to estimate θ via a multistage sampling plan with the following structure. The sampling process is divided into s stages, where s can be infinity or a positive integer. The continuation or termination of sampling is determined by decision variables. For each stage with index ℓ , a decision variable $\mathbf{D}_\ell = \mathcal{D}_\ell(X_1, \dots, X_{\mathbf{n}_\ell})$ is defined based on samples $X_1, \dots, X_{\mathbf{n}_\ell}$, where \mathbf{n}_ℓ is the number of samples available at the ℓ -th stage. It should be noted that \mathbf{n}_ℓ can be a random number, depending on specific sampling schemes. The decision variable \mathbf{D}_ℓ assumes only two possible values 0, 1 with the notion that the sampling is continued until $\mathbf{D}_\ell = 1$ for some ℓ . For the ℓ -th stage, an estimator $\hat{\theta}_\ell$ for θ is defined based on samples $X_1, \dots, X_{\mathbf{n}_\ell}$. Let \mathbf{l} denote the index of stage when the sampling is terminated. Then, the point estimator for θ , denoted by $\hat{\theta}$, is equal to $\hat{\theta}_{\mathbf{l}}$. The decision variables \mathbf{D}_ℓ can be defined in terms of estimators $\hat{\theta}_\ell$ and confidence intervals (L_ℓ, U_ℓ) , where the lower confidence limit L_ℓ and upper confidence limit U_ℓ are functions of $X_1, \dots, X_{\mathbf{n}_\ell}$ for $\ell = 1, \dots, s$. Depending on various error criterion, we have different sampling plans as follows.

Theorem 2 *Let $\varepsilon > 0$, $\zeta > 0$ and $\delta \in (0, 1)$. For $\ell = 1, \dots, s$, let (L_ℓ, U_ℓ) be a confidence interval such that $\Pr\{L_\ell < \theta < U_\ell\} > 1 - \zeta\delta$. Suppose the stopping rule is that sampling is continued until*

$U_\ell - \varepsilon < \widehat{\theta}_\ell < L_\ell + \varepsilon$ at some stage with index ℓ . Then, $\Pr\{|\widehat{\theta} - \theta| < \varepsilon\} > 1 - \delta$ provided that $s\zeta < 1$ and that $\Pr\{U_s - \varepsilon < \widehat{\theta}_s < L_s + \varepsilon\} = 1$.

We would like to note that, for estimating the mean value of a random variable bounded in $[a, b]$, Theorem 2 can be applied based on the following choice:

- (i) The sample sizes of the sampling plan are chosen as deterministic integers $n_1 < \dots < n_s$ such that $n_s > \frac{(b-a)^2}{2\varepsilon^2} \ln \frac{2s}{\delta}$.
- (ii) The confidence intervals are constructed by virtue of Theorem 1.

Theorem 3 Let $\varepsilon > 0$, $\zeta > 0$ and $\delta \in (0, 1)$. For $\ell = 1, \dots, s$, let (L_ℓ, U_ℓ) be a confidence interval such that $\Pr\{L_\ell < \theta < U_\ell\} > 1 - \zeta\delta$. Suppose the stopping rule is that sampling is continued until $[1 - \text{sgn}(\widehat{\theta}_\ell) \varepsilon]U_\ell < \widehat{\theta}_\ell < [1 + \text{sgn}(\widehat{\theta}_\ell) \varepsilon]L_\ell$ at some stage with index ℓ . Then, $\Pr\{|\widehat{\theta} - \theta| < \varepsilon|\theta|\} > 1 - \delta$ provided that $s\zeta < 1$ and that $\Pr\{[1 - \text{sgn}(\widehat{\theta}_s) \varepsilon]U_s < \widehat{\theta}_s < [1 + \text{sgn}(\widehat{\theta}_s) \varepsilon]L_s\} = 1$, where $\text{sgn}(x)$ is the sign function which assumes values 1, 0 and -1 for $x > 0$, $x = 0$ and $x < 0$ respectively.

We would like to note that, for estimating the mean value of a random variable bounded in $[0, 1]$, we can use Theorems 1 and 3 based on multistage inverse sampling.

Theorem 4 Let $0 < \delta < 1$, $\varepsilon_a > 0$, $\varepsilon_r > 0$ and $\zeta > 0$. For $\ell = 1, \dots, s$, let (L_ℓ, U_ℓ) be a confidence interval such that $\Pr\{L_\ell < \theta < U_\ell\} > 1 - \zeta\delta$. Suppose the stopping rule is that sampling is continued until $U_\ell - \max(\varepsilon_a, \text{sgn}(\widehat{\theta}_\ell) \varepsilon_r U_\ell) < \widehat{\theta}_\ell < L_\ell + \max(\varepsilon_a, \text{sgn}(\widehat{\theta}_\ell) \varepsilon_r L_\ell)$ at some stage with index ℓ . Then, $\Pr\left\{\left|\widehat{\theta} - \theta\right| < \varepsilon_a \text{ or } \left|\widehat{\theta} - \theta\right| < \varepsilon_r |\theta|\right\} \geq 1 - \delta$ provided that $s\zeta < 1$ and that $\Pr\{U_s - \max(\varepsilon_a, \text{sgn}(\widehat{\theta}_s) \varepsilon_r U_s) < \widehat{\theta}_s < L_s + \max(\varepsilon_a, \text{sgn}(\widehat{\theta}_s) \varepsilon_r L_s)\} = 1$.

For estimating the mean value of a random variable bounded in $[a, b]$, Theorem 4 can be applied based on the following choice:

- (i) The sample sizes of the sampling plan are chosen as deterministic integers $n_1 < \dots < n_s$ such that $n_s > \frac{(b-a)^2}{2\varepsilon^2} \ln \frac{2s}{\delta}$.
- (ii) The confidence intervals are constructed by virtue of Theorem 1.

In Theorems 2–4, the number of stages, s , is assumed to be a finite integer. In some situations, a sampling plan with a finite number of stages is impossible to guarantee the prescribed precision and confidence level. In this regard, the following theorems are useful.

Theorem 5 Let $\varepsilon > 0$, $\zeta > 0$ and $\delta \in (0, 1)$. Let τ be a positive integer. Let (L_ℓ, U_ℓ) be a confidence interval such that $\Pr\{L_\ell < \theta < U_\ell\} > 1 - \zeta\delta$ for $\ell \leq \tau$ and that $\Pr\{L_\ell < \theta < U_\ell\} > 1 - \zeta\delta 2^{\tau-\ell}$ for $\ell > \tau$. Suppose the stopping rule is that sampling is continued until $U_\ell - \varepsilon < \widehat{\theta}_\ell < L_\ell + \varepsilon$ at some stage with index ℓ . Then, $\Pr\{|\widehat{\theta} - \theta| < \varepsilon\} > 1 - \delta$ provided that $(\tau + 1)\zeta < 1$ and that $\Pr\{\mathbf{l} < \infty\} = 1$.

Theorem 6 Let $\varepsilon > 0$, $\zeta > 0$ and $\delta \in (0, 1)$. Let τ be a positive integer. Let (L_ℓ, U_ℓ) be a confidence interval such that $\Pr\{L_\ell < \theta < U_\ell\} > 1 - \zeta\delta$ for $\ell \leq \tau$ and that $\Pr\{L_\ell < \theta < U_\ell\} > 1 - \zeta\delta 2^{\tau-\ell}$ for $\ell > \tau$. Suppose the stopping rule is that sampling is continued until $[1 - \text{sgn}(\hat{\theta}_\ell) \varepsilon] U_\ell < \hat{\theta}_\ell < [1 + \text{sgn}(\hat{\theta}_\ell) \varepsilon] L_\ell$ at some stage with index ℓ . Then, $\Pr\{|\hat{\theta} - \theta| < \varepsilon|\theta|\} > 1 - \delta$ provided that $(\tau + 1)\zeta < 1$ and that $\Pr\{\mathbf{l} < \infty\} = 1$.

Theorem 7 Let $0 < \delta < 1$, $\varepsilon_a > 0$, $\varepsilon_r > 0$ and $\zeta > 0$. Let τ be a positive integer. Let (L_ℓ, U_ℓ) be a confidence interval such that $\Pr\{L_\ell < \theta < U_\ell\} > 1 - \zeta\delta$ for $\ell \leq \tau$ and that $\Pr\{L_\ell < \theta < U_\ell\} > 1 - \zeta\delta 2^{\tau-\ell}$ for $\ell > \tau$. Suppose the stopping rule is that sampling is continued until $U_\ell - \max(\varepsilon_a, \text{sgn}(\hat{\theta}_\ell) \varepsilon_r U_\ell) < \hat{\theta}_\ell < L_\ell + \max(\varepsilon_a, \text{sgn}(\hat{\theta}_\ell) \varepsilon_r L_\ell)$ at some stage with index ℓ . Then, $\Pr\left\{|\hat{\theta} - \theta| < \varepsilon_a \text{ or } |\hat{\theta} - \theta| < \varepsilon_r|\theta|\right\} \geq 1 - \delta$ provided that $(\tau + 1)\zeta < 1$ and that $\Pr\{\mathbf{l} < \infty\} = 1$.

We would like to note that, for estimating the mean value of a random variable bounded in $[a, b]$, Theorems 5–7 can be used since it can be shown that $\Pr\{\mathbf{l} < \infty\} = 1$ as a consequence of using the confidence interval described by Theorem 1.

References

- [1] Hoeffding, W. (1963), “Probability inequalities for sums of bounded random variables”, *J. Amer. Statist. Assoc.*, vol. 58, pp. 13-29.
- [2] Massart, P. (1990), “The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality”, *The Annals of Probability*, vol. 18, pp. 1269-1283.